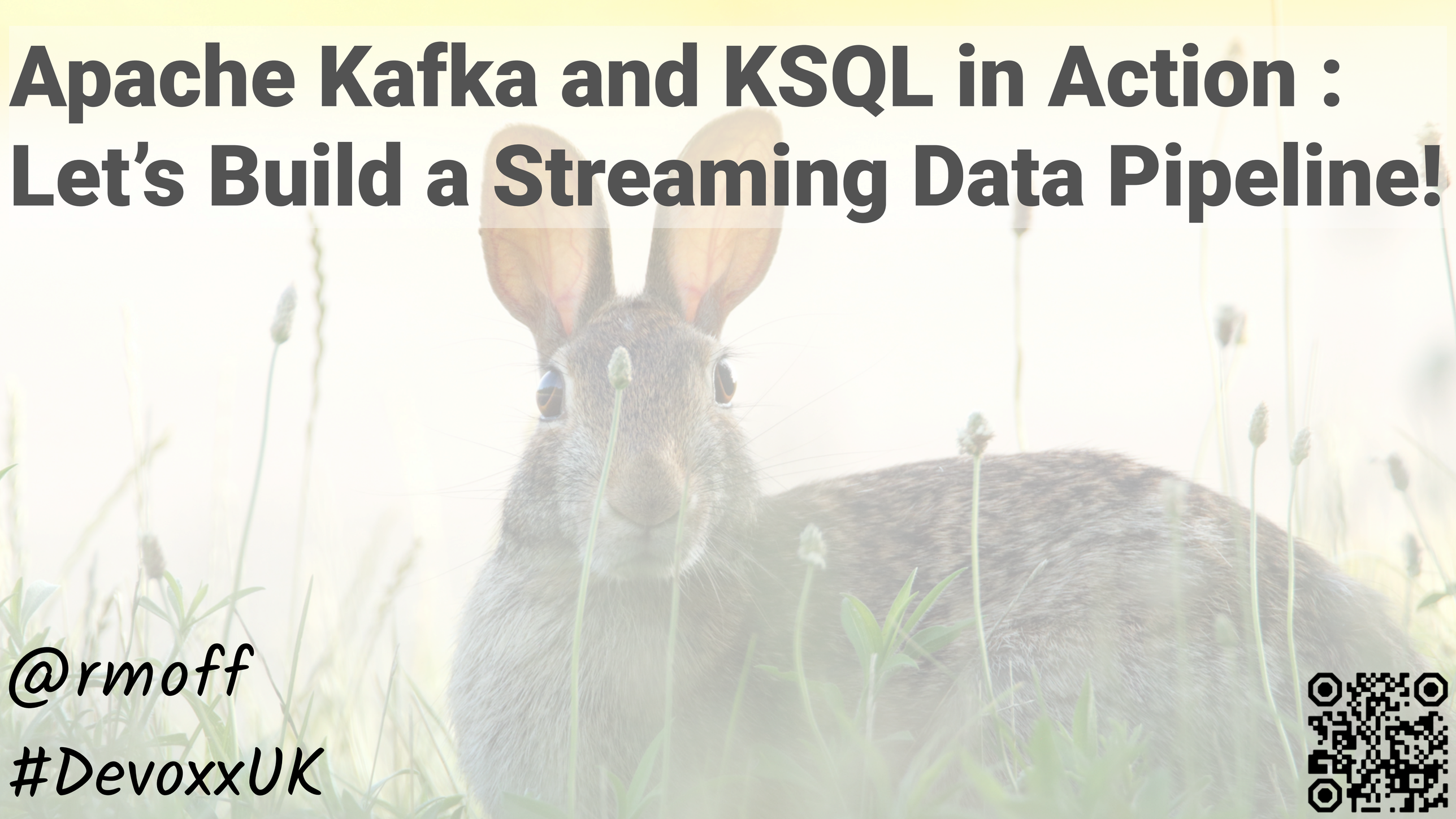
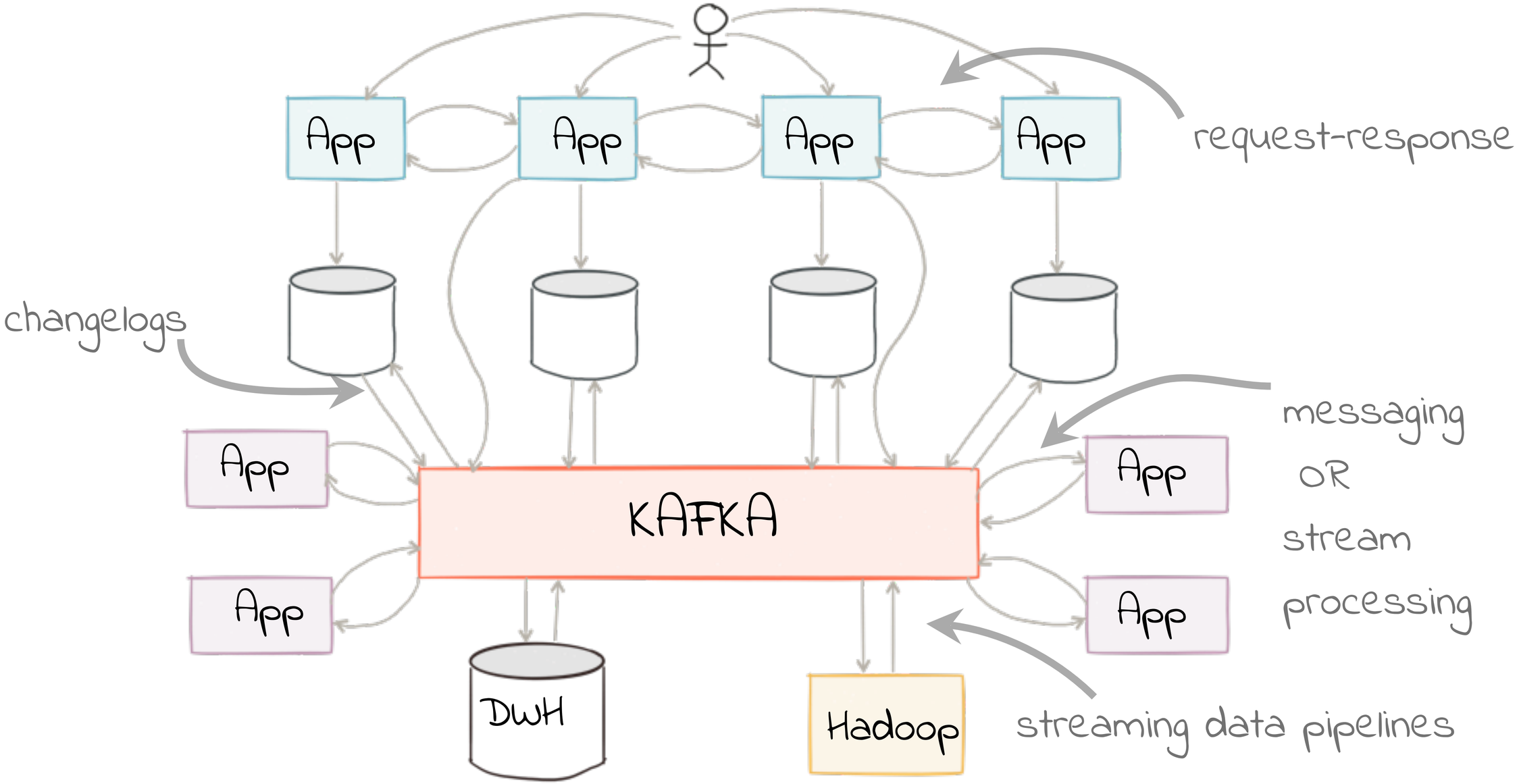
Apache Kafka and KSQL in Action : Let's Build a Streaming Data Pipeline!



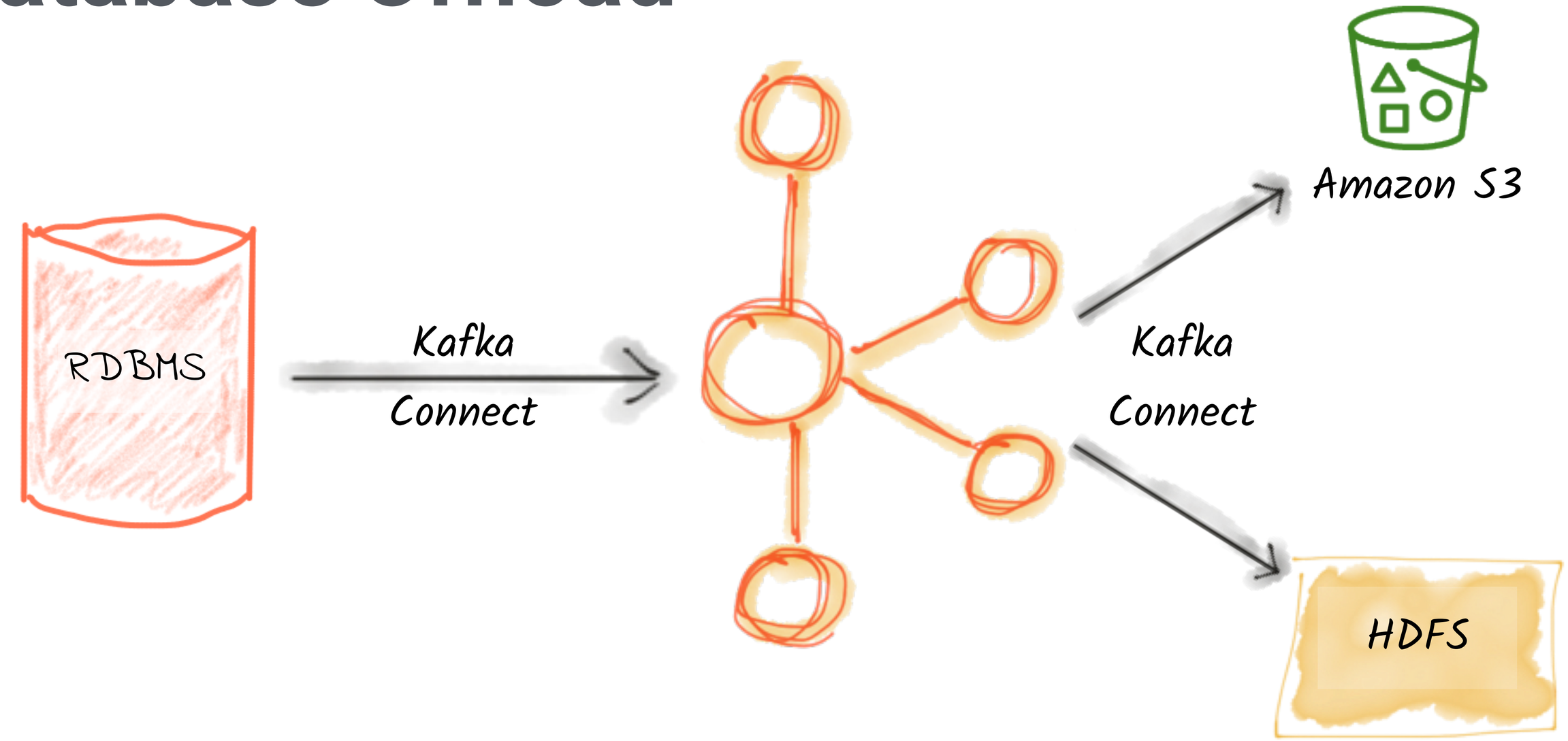
@rmoff

#DevoxxUK

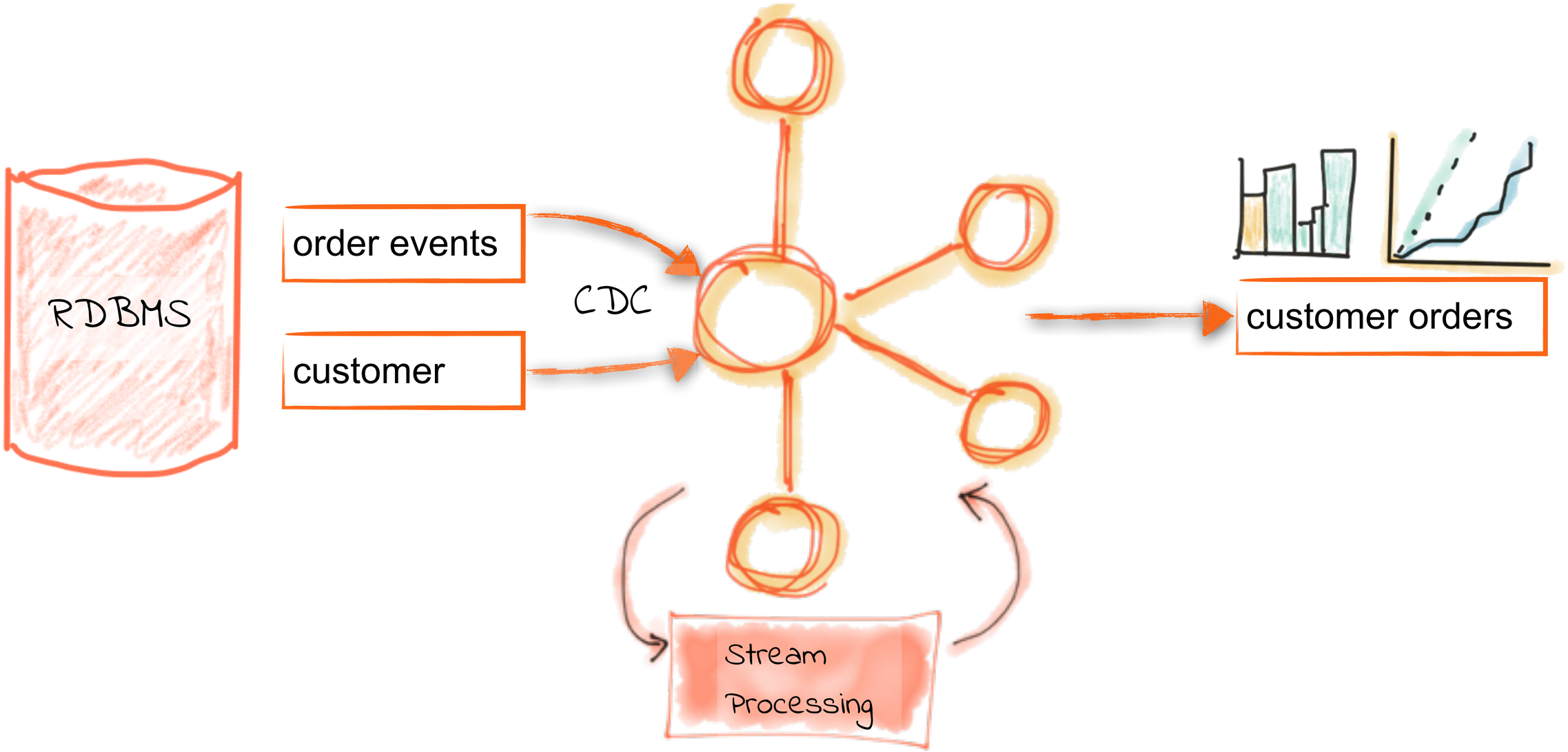
Kafka is a Streaming Platform



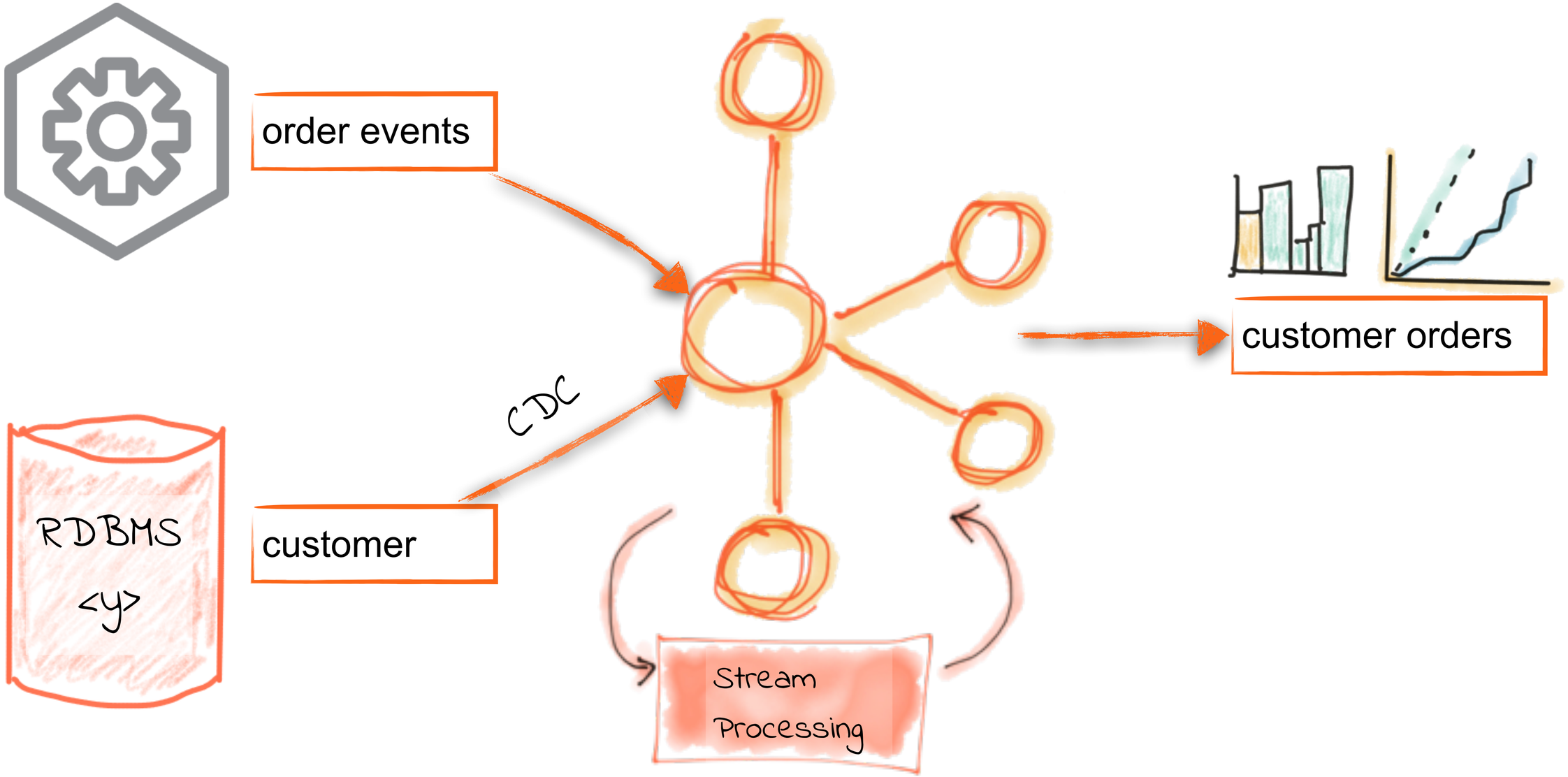
Database Offload



Stream Processing with Apache Kafka and KSQL

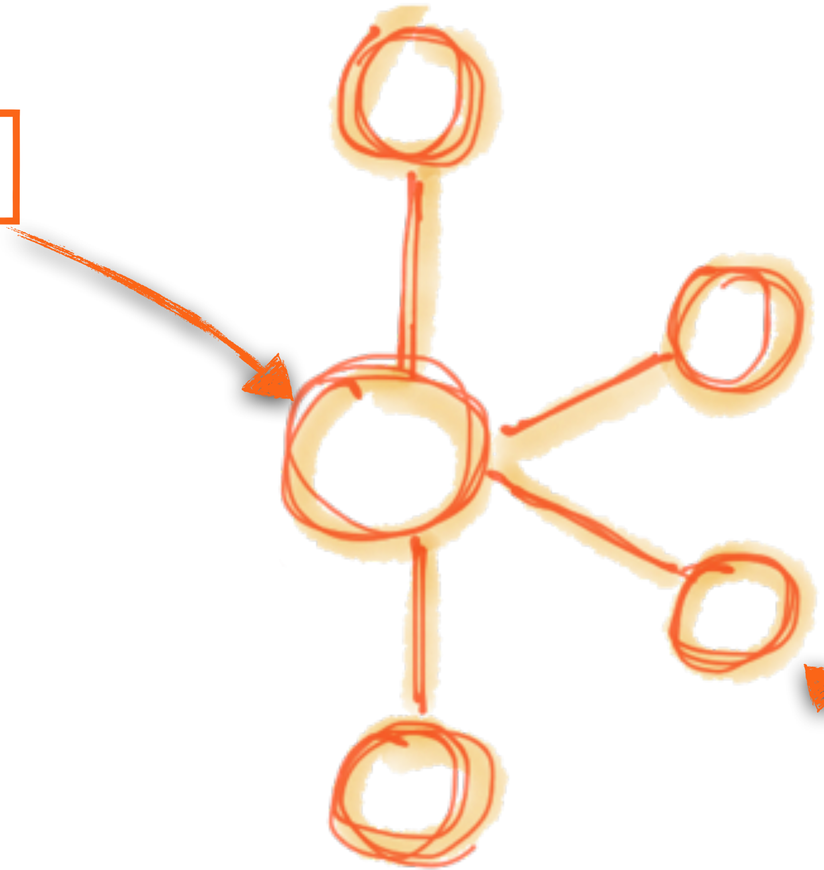


Real-time Event Stream Enrichment



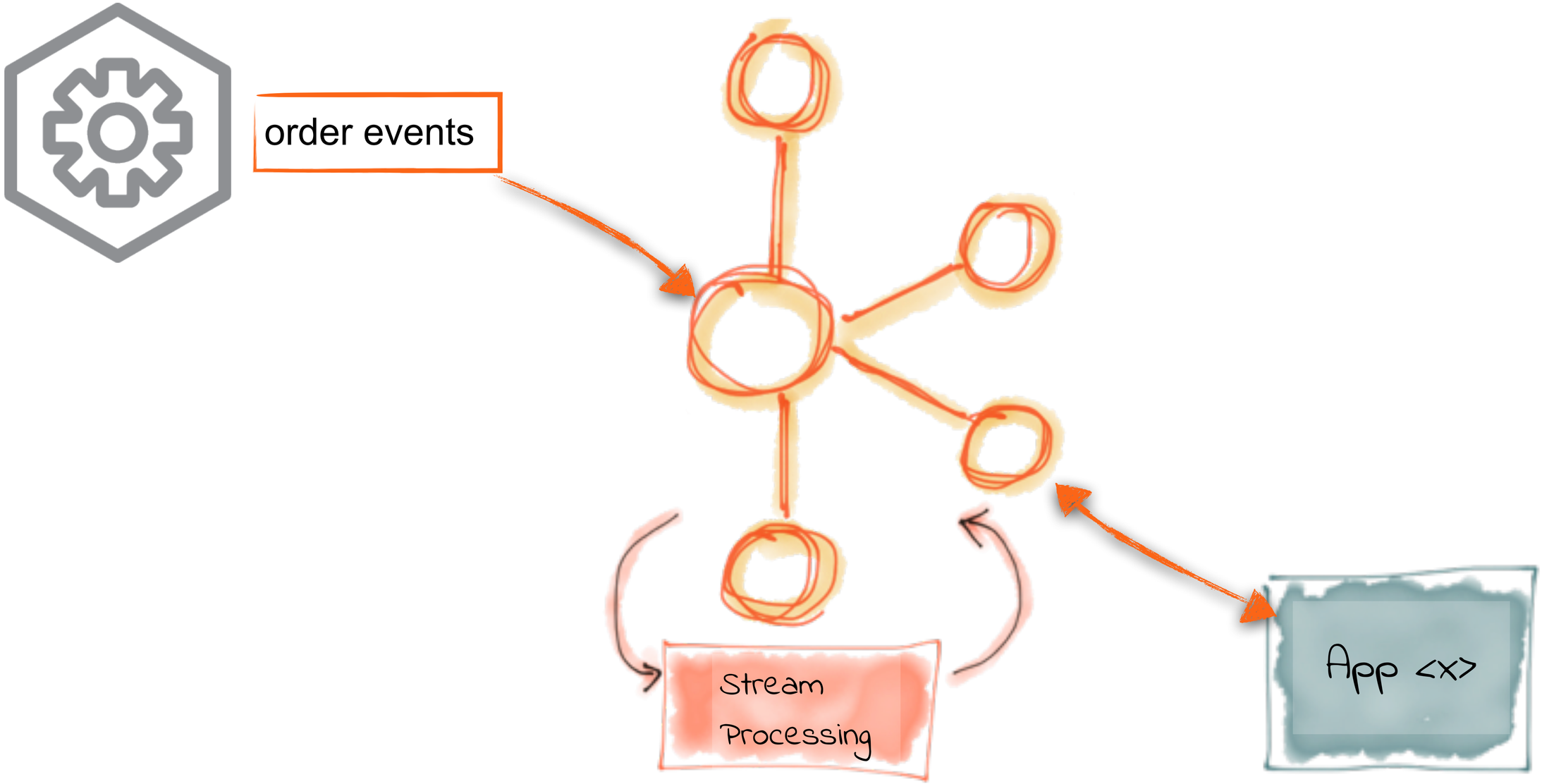
Message Bus

order events



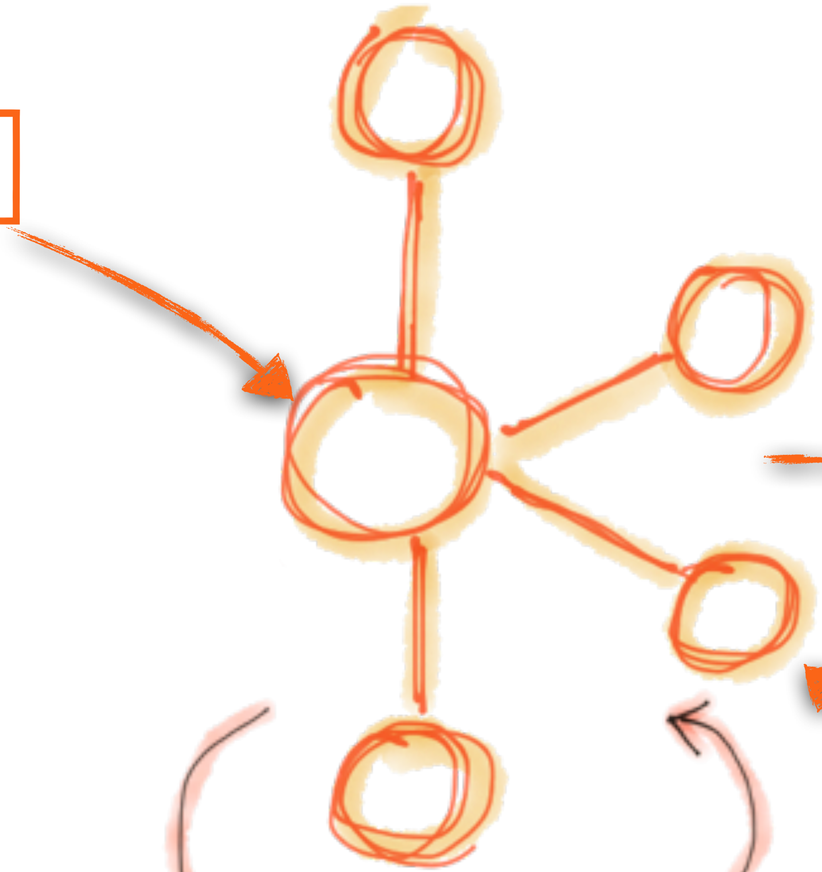
App <x>

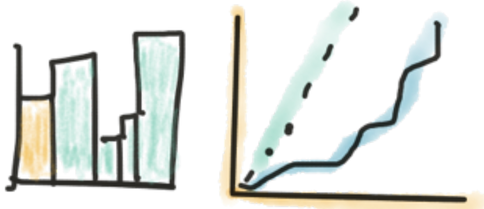
Stream Processing



Transform Once, Use Many

order events



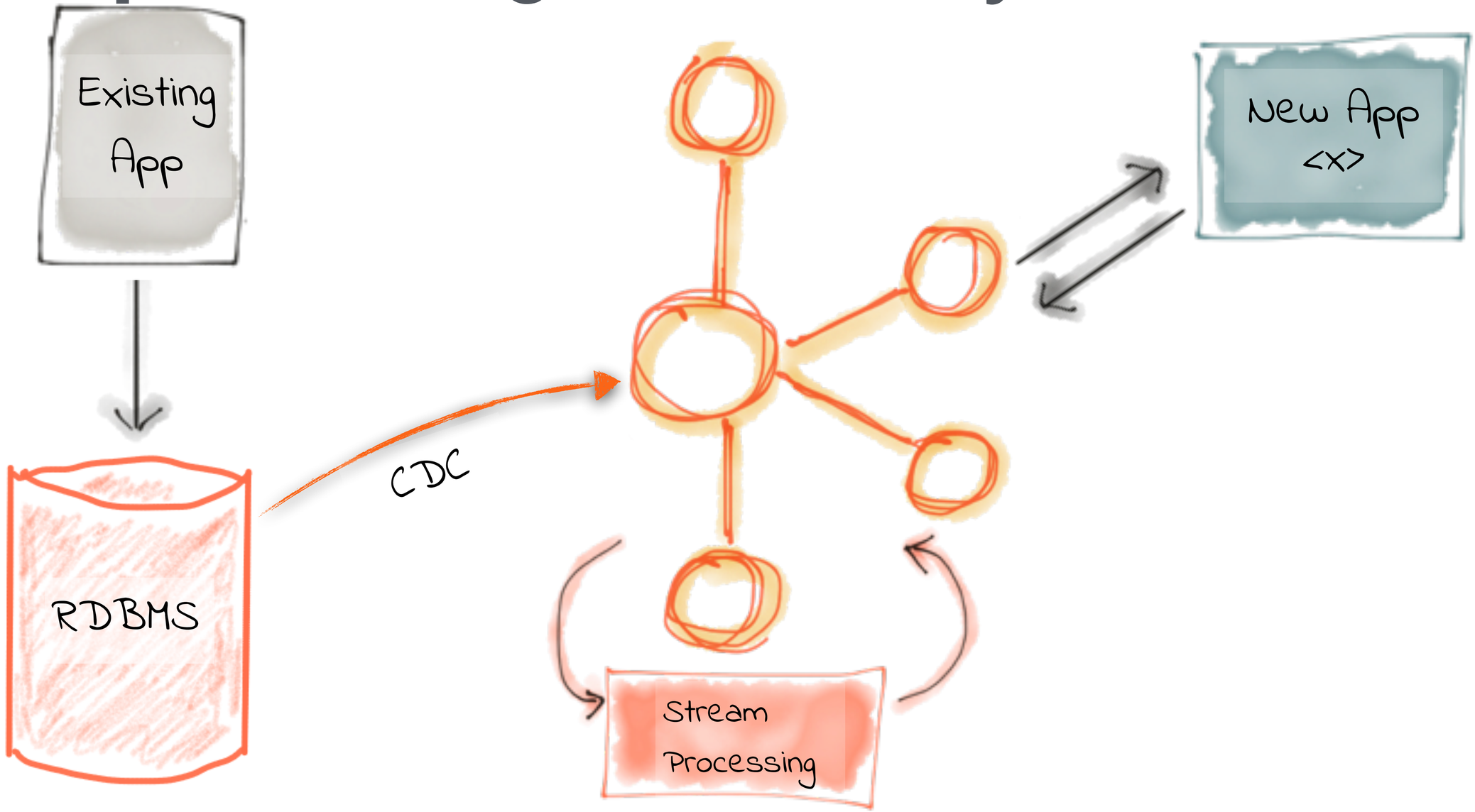


Completed orders

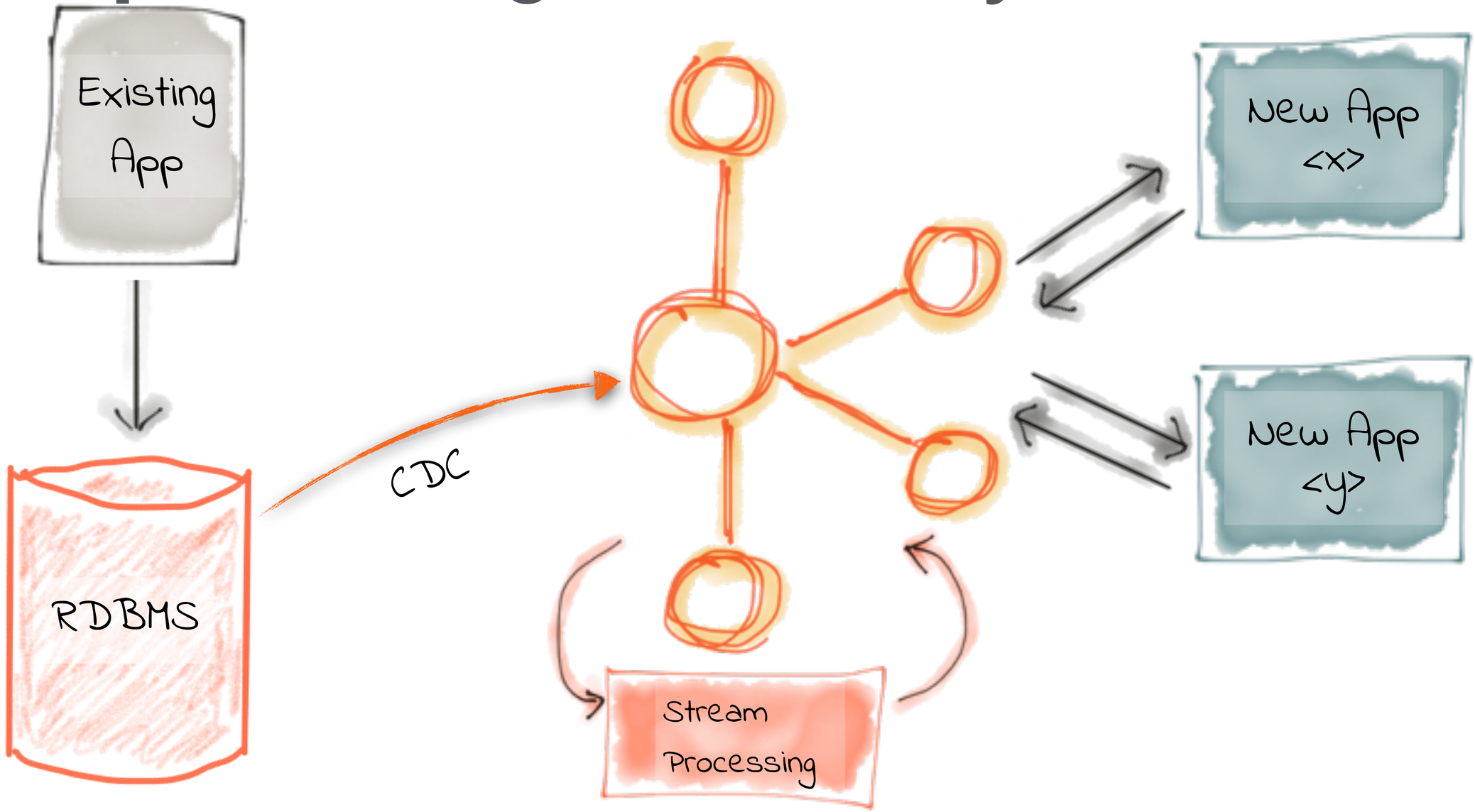
Stream Processing

App <x>

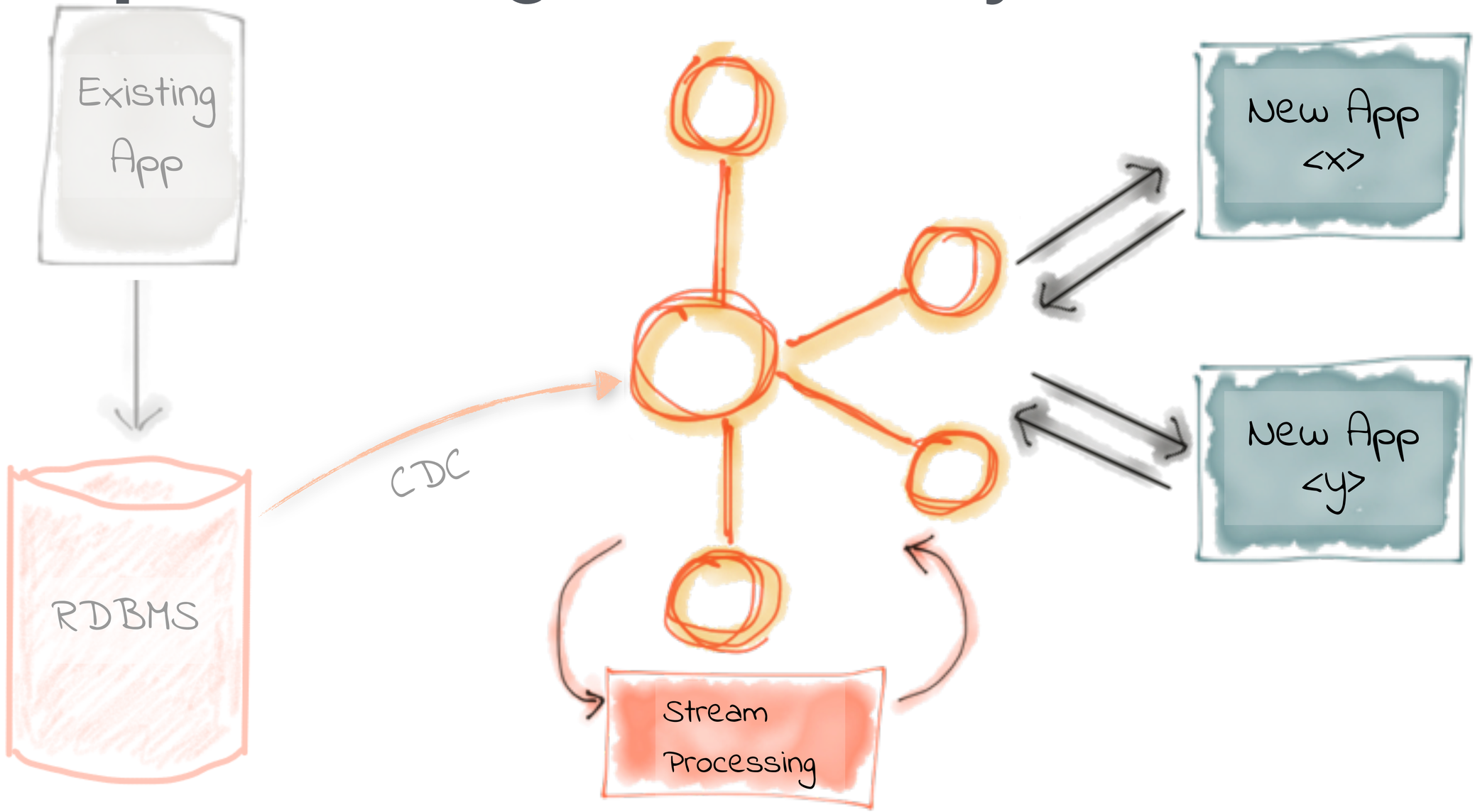
Evolve processing from old systems to new



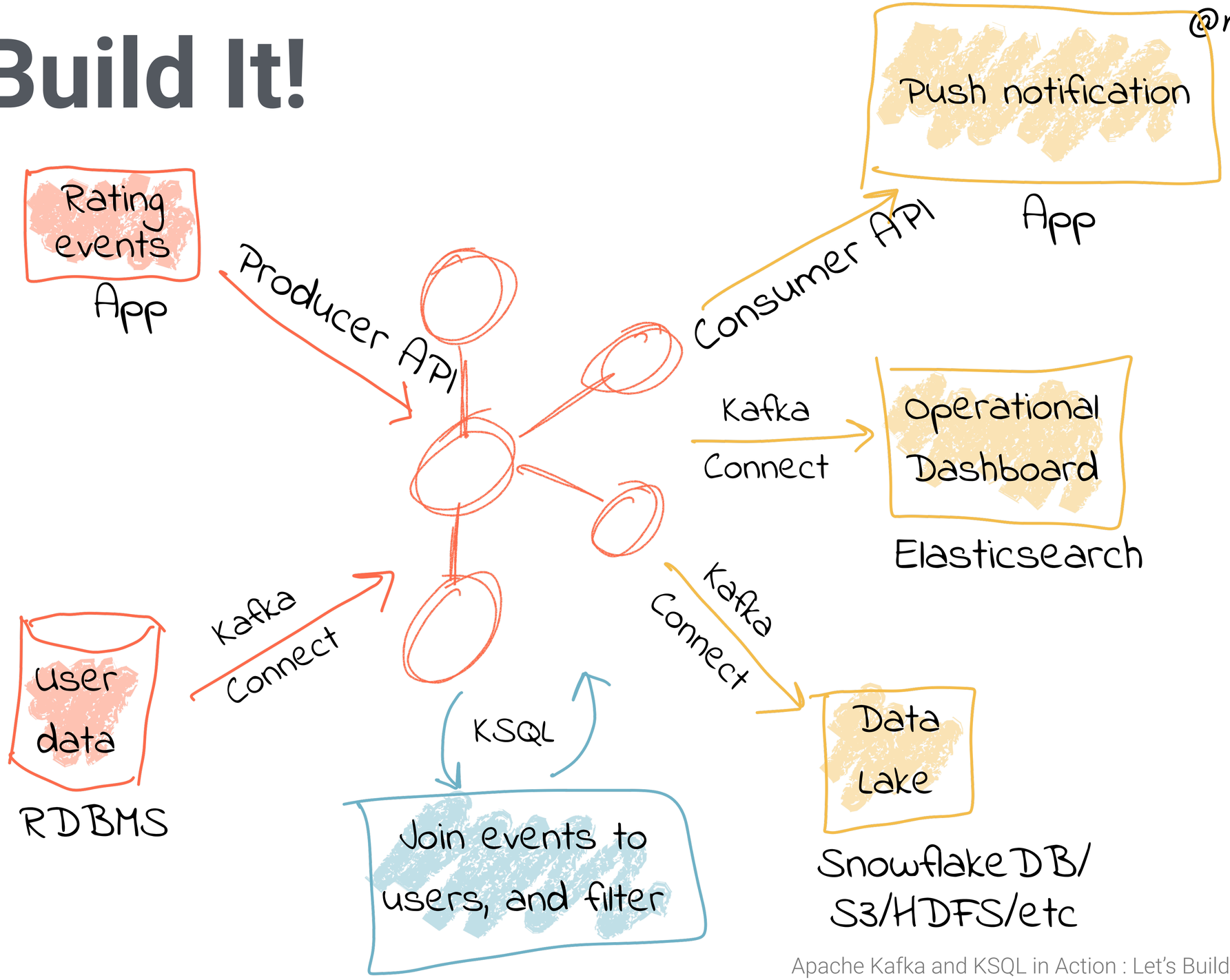
Evolve processing from old systems to new



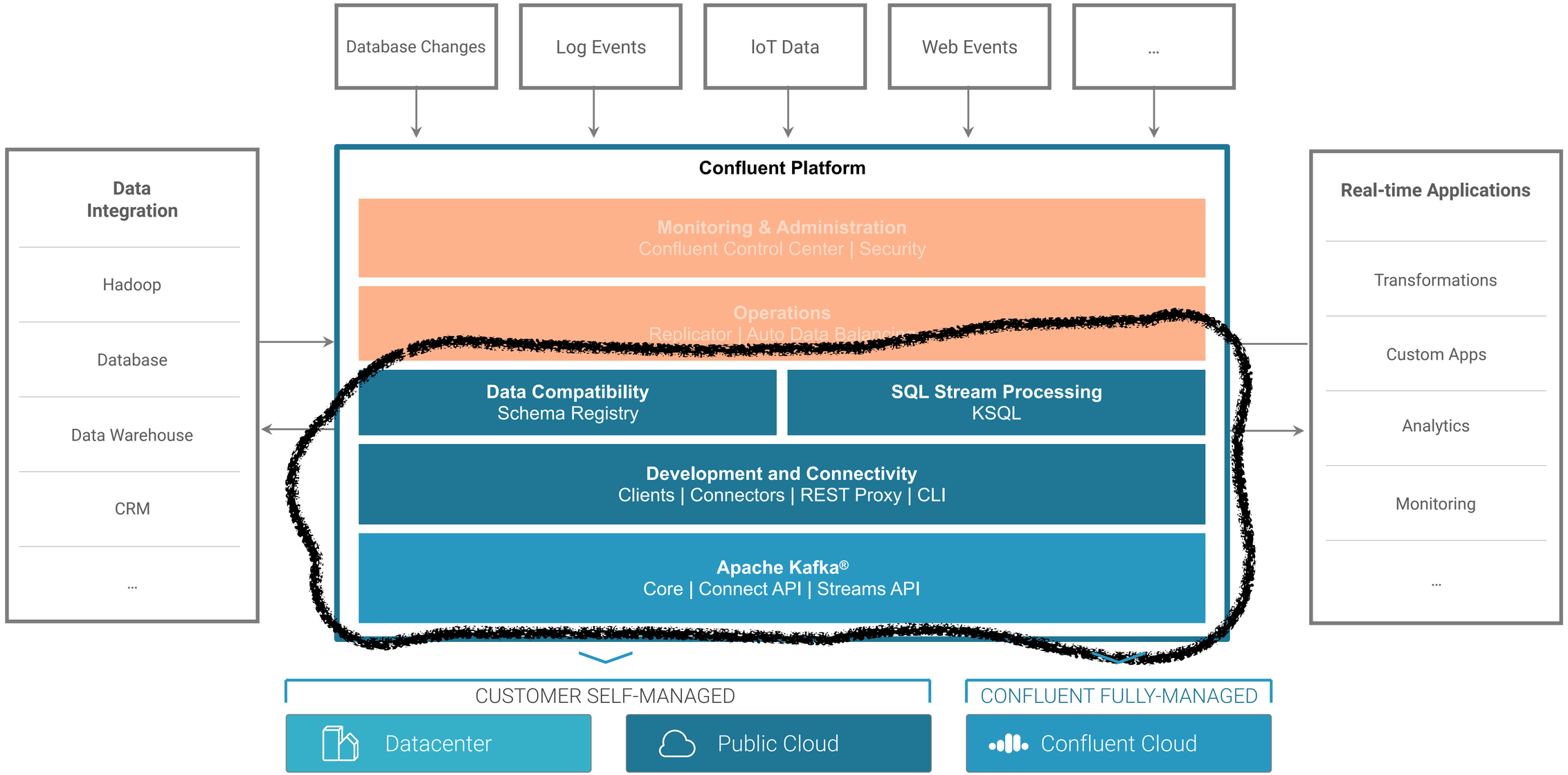
Evolve processing from old systems to new



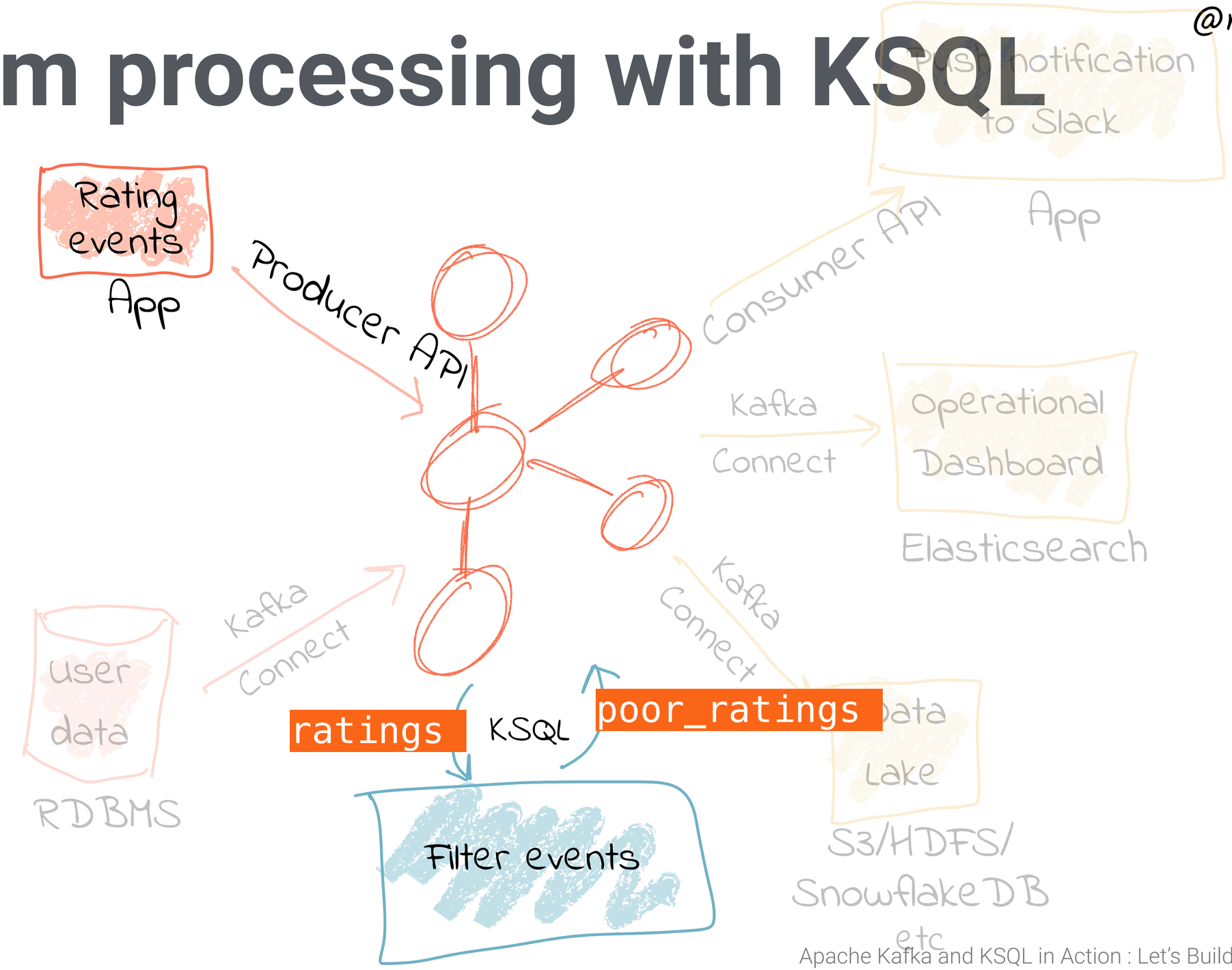
Let's Build It!



Confluent Community Components
Apache Kafka with a bunch of cool stuff! For free!

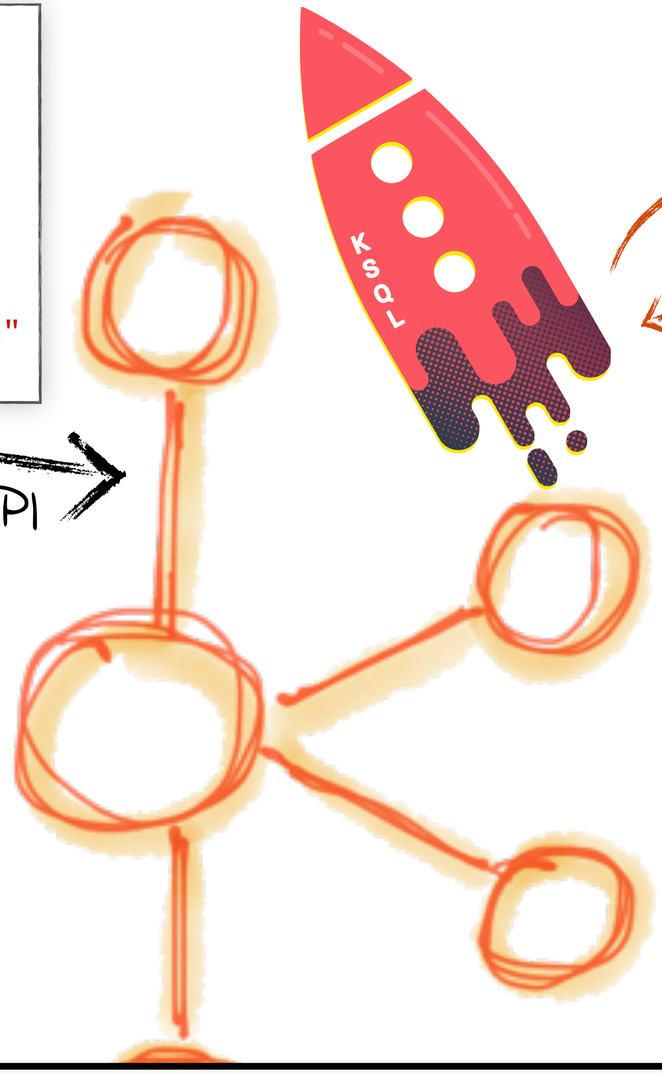


Stream processing with KSQL



```
{
 "rating_id": 5313,
 "user_id": 3,
 "stars": 4,
 "route_id": 6975,
 "rating_time": 1519304105213,
 "channel": "web",
 "message": "worst. flight. ever. #neveragain"
}
```

Producer API

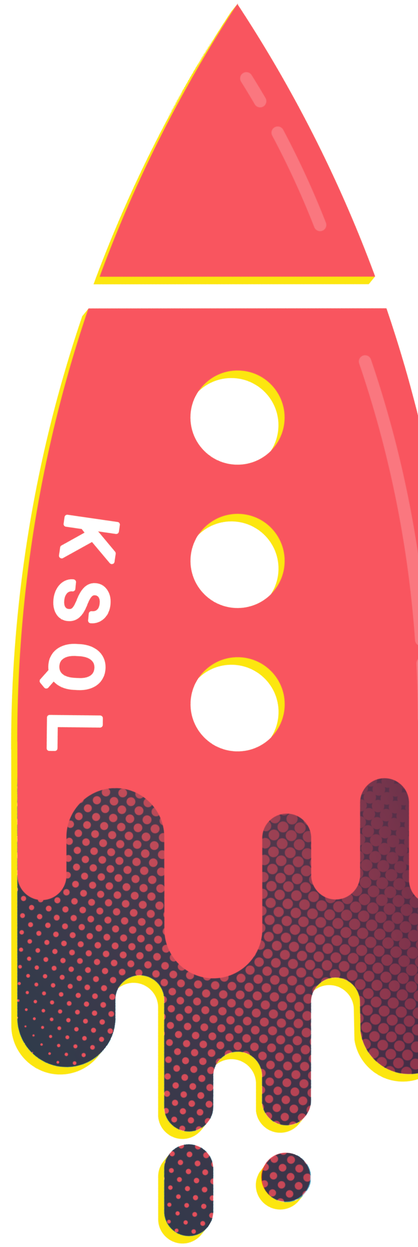


Filter all ratings where STARS<3

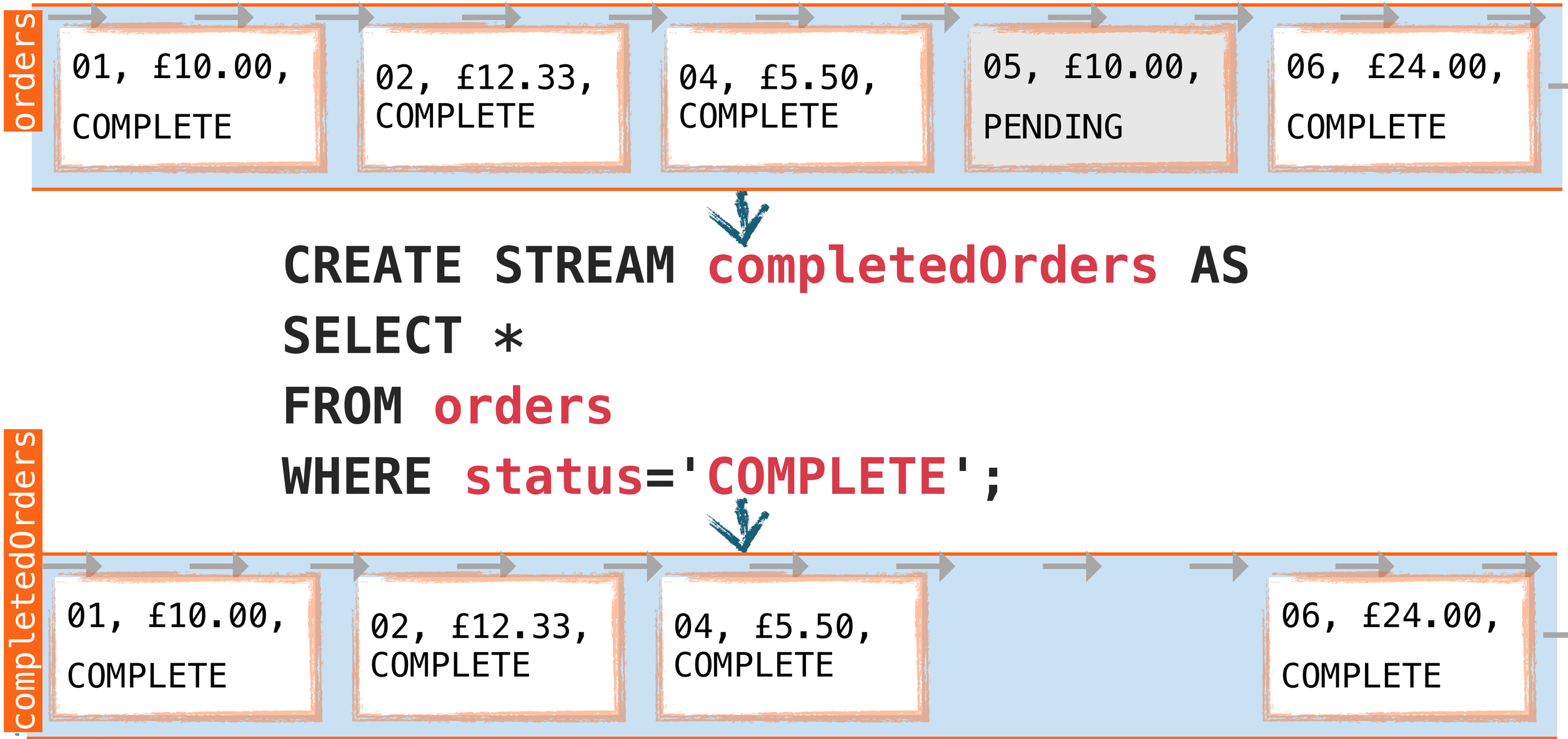
POOR_RATINGS

```
CREATE STREAM POOR_RATINGS AS
SELECT * FROM ratings WHERE STARS <3
```

KSQL
is the
Streaming
SQL Engine
for
Apache Kafka



Filter messages with KSQL



Drop columns with KSQL

customer

```
{"id":1,
"name":"Dana Lidgerton",
"card":"5048370182840140}
```

```
{"id":2,
"name":"Milo Wellsman",
"card":"3557977885537506}
```

```
{"id":3,
"name":"Dolph Cleeton",
"card":"3586303633007251}
```

```
CREATE STREAM customerNoCC AS
SELECT ID, NAME
FROM customer;
```

customerNoCC

```
{"id":1,
"name":"Dana Lidgerton"}
```

```
{"id":2,
"name":"Milo Wellsman"}
```

```
{"id":3,
"name":"Dolph Cleeton"}
```

Stateful aggregation with KSQL

customer

```
{"id":1,
"name":"Dana Lidgerton",
"country":"UK"}
```

```
{"id":2,
"name":"Milo Wellsman",
"country":"UK"}
```

```
{"id":3,
"name":"Dolph Cleeton",
"country":"Germany"}
```

```
CREATE STREAM customersByCountry AS
SELECT country, COUNT(*) AS customerCount
FROM customer WINDOW TUMBLING (SIZE 1 HOUR)
     GROUP BY country;
```

customersByCountry

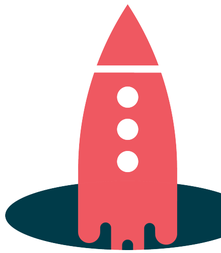
```
{"country":"UK",
"customerCount":2}
```

```
{"country":"Germany",
"customerCount":1}
```

KSQL for Anomaly Detection

Identifying patterns or anomalies in real-time data, surfaced in milliseconds

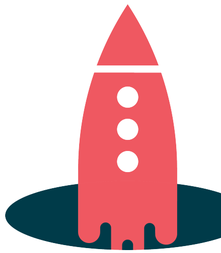
```
CREATE TABLE possible_fraud AS
  SELECT card_number, count(*)
    FROM authorization_attempts
    WINDOW TUMBLING (SIZE 5 SECONDS)
    GROUP BY card_number
    HAVING count(*) > 3;
```



KSQL for Data Transformation

Make simple derivations of existing topics from the command line

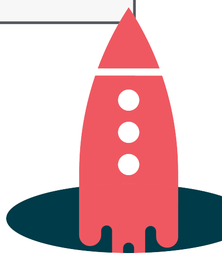
```
CREATE STREAM pageviews
  WITH (PARTITIONS=4,
        VALUE_FORMAT='AVRO') AS
  SELECT * FROM pageviews_json;
```



KSQL for Streaming ETL

Joining, filtering, and aggregating streams of event data

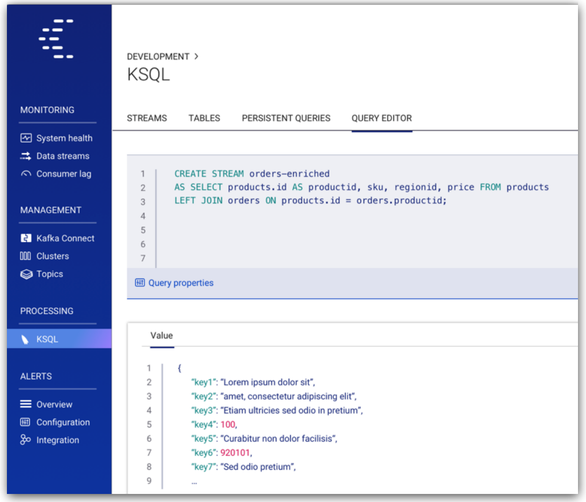
```
CREATE STREAM vip_actions AS
  SELECT userid, page, action
  FROM clickstream c
  LEFT JOIN users u
    ON c.userid = u.user_id
  WHERE u.level = 'Platinum';
```



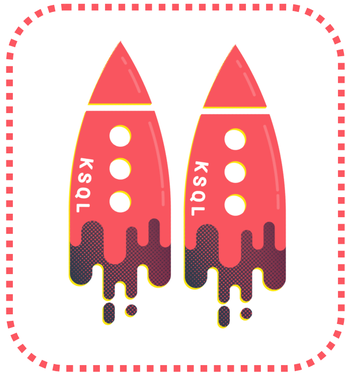
KSQL in Development and Production

Interactive KSQL
for development and testing

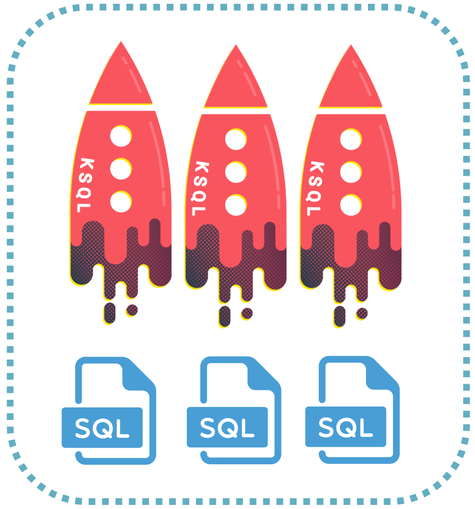
Headless KSQL
for Production



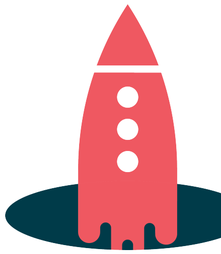
REST



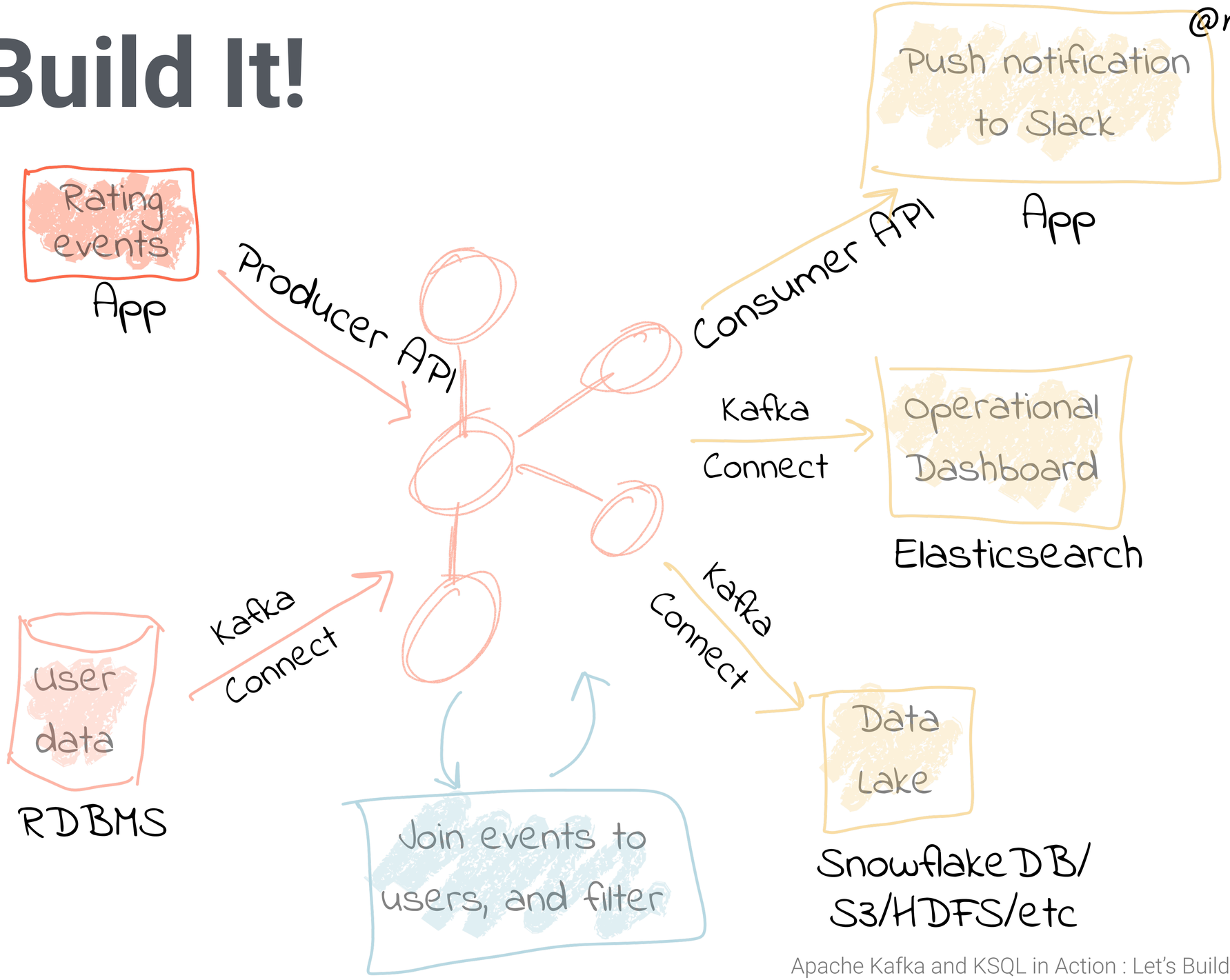
Desired KSQL queries have been identified



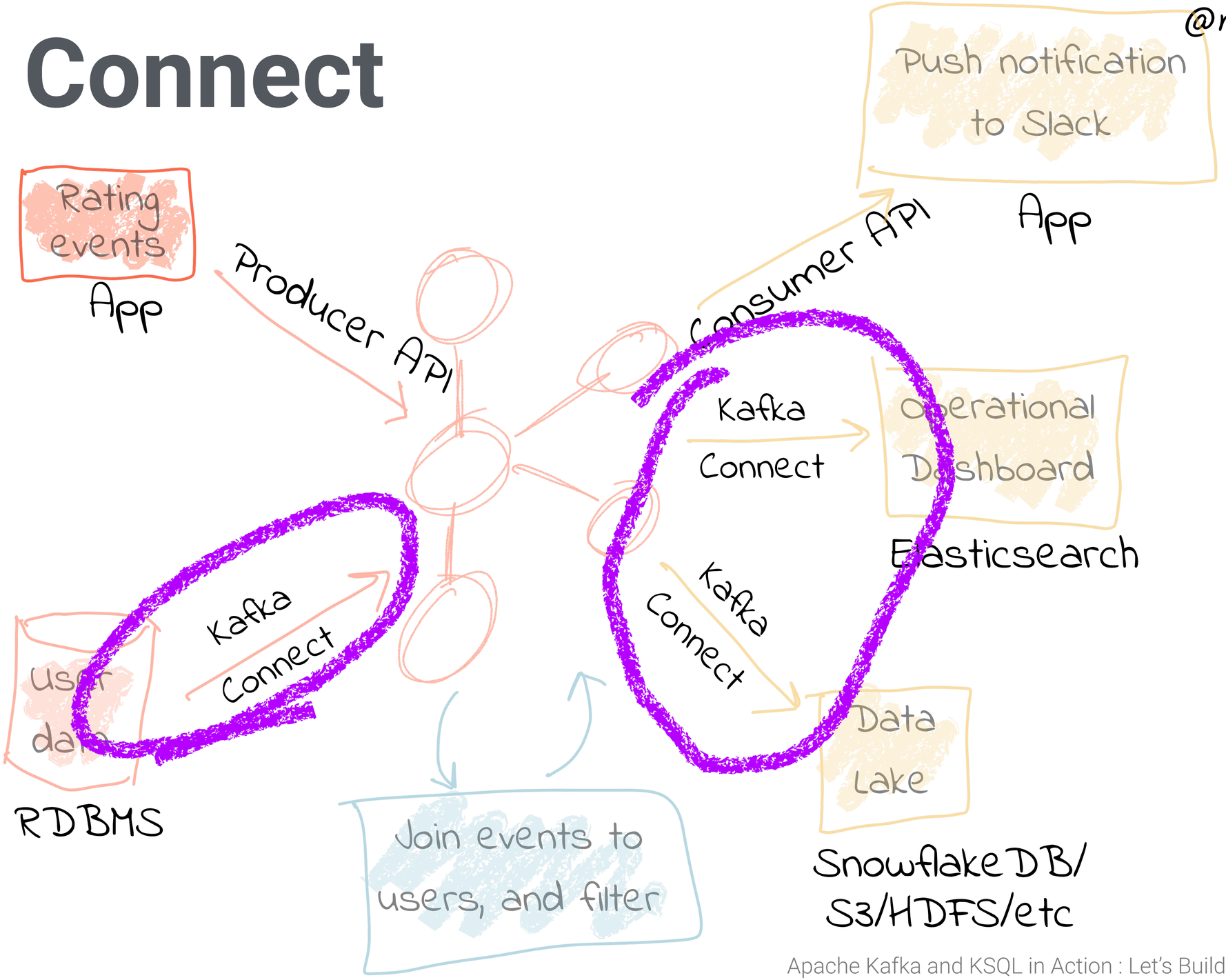
"Hmm, let me try out this idea..."



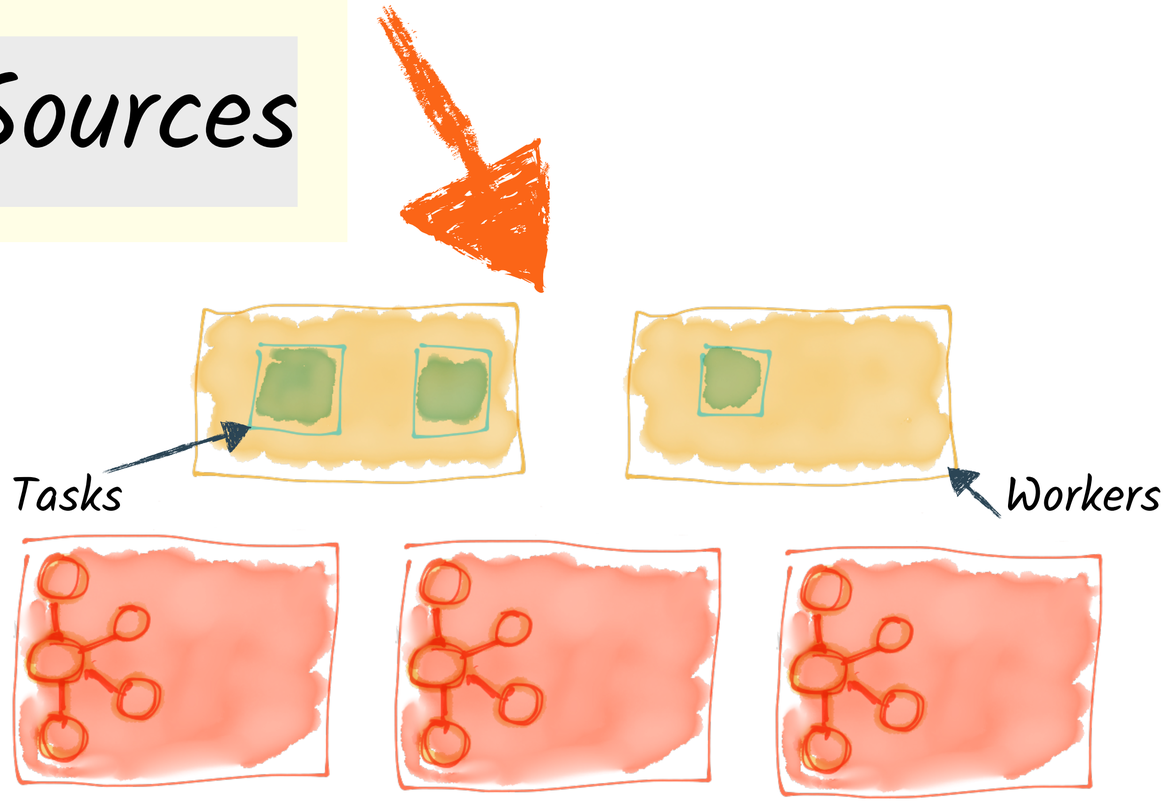
Let's Build It!



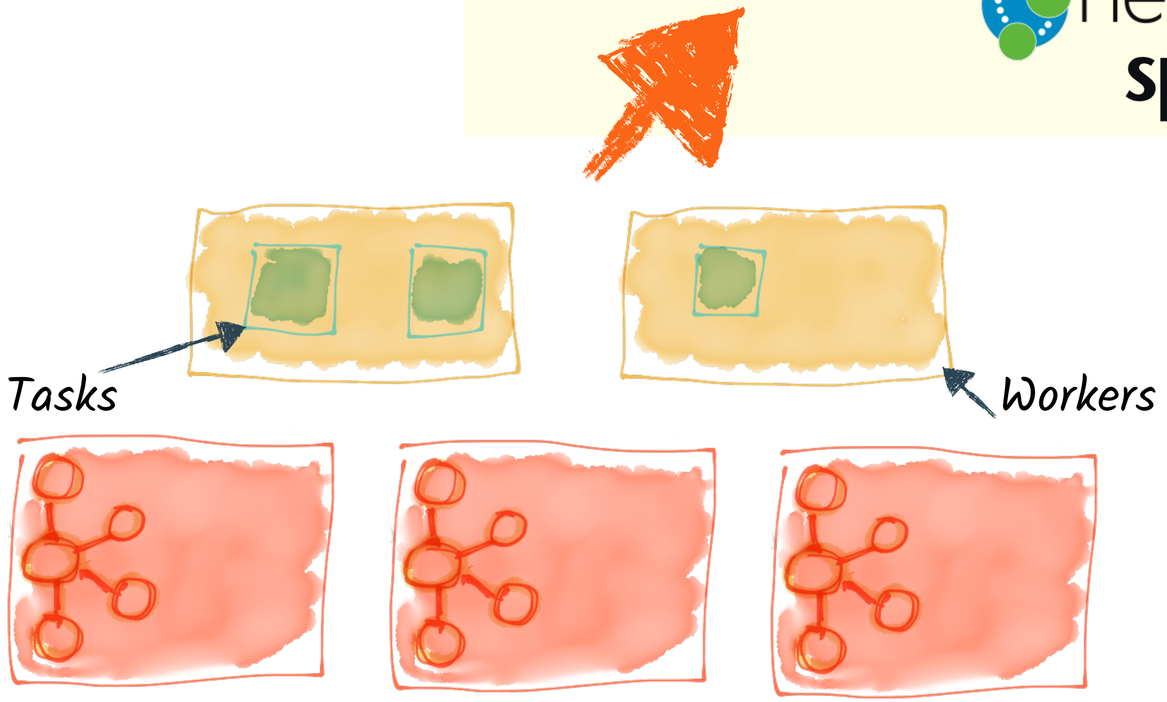
Kafka Connect



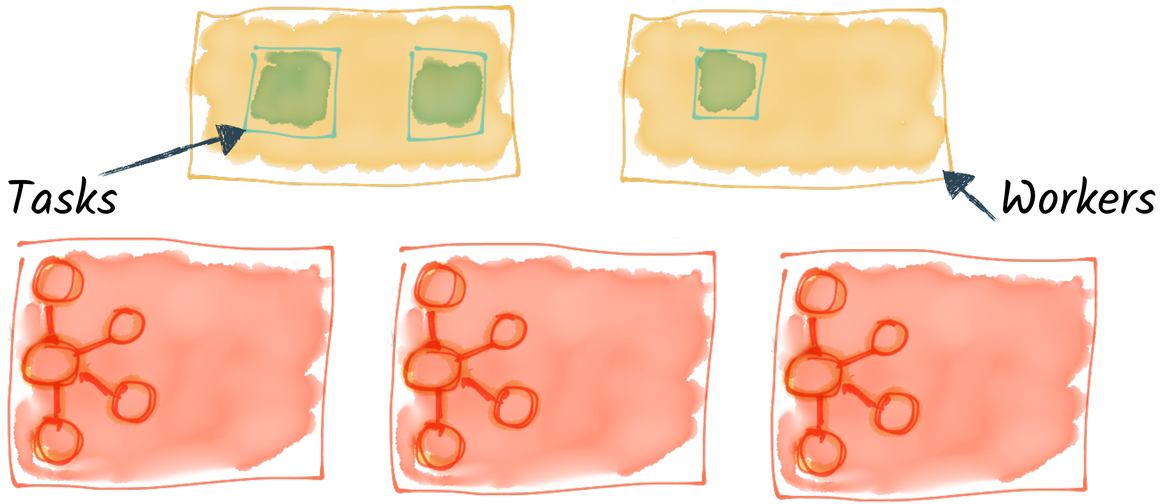
Streaming Integration with Kafka Connect



Streaming Integration with Kafka Connect



Streaming Integration with Kafka Connect



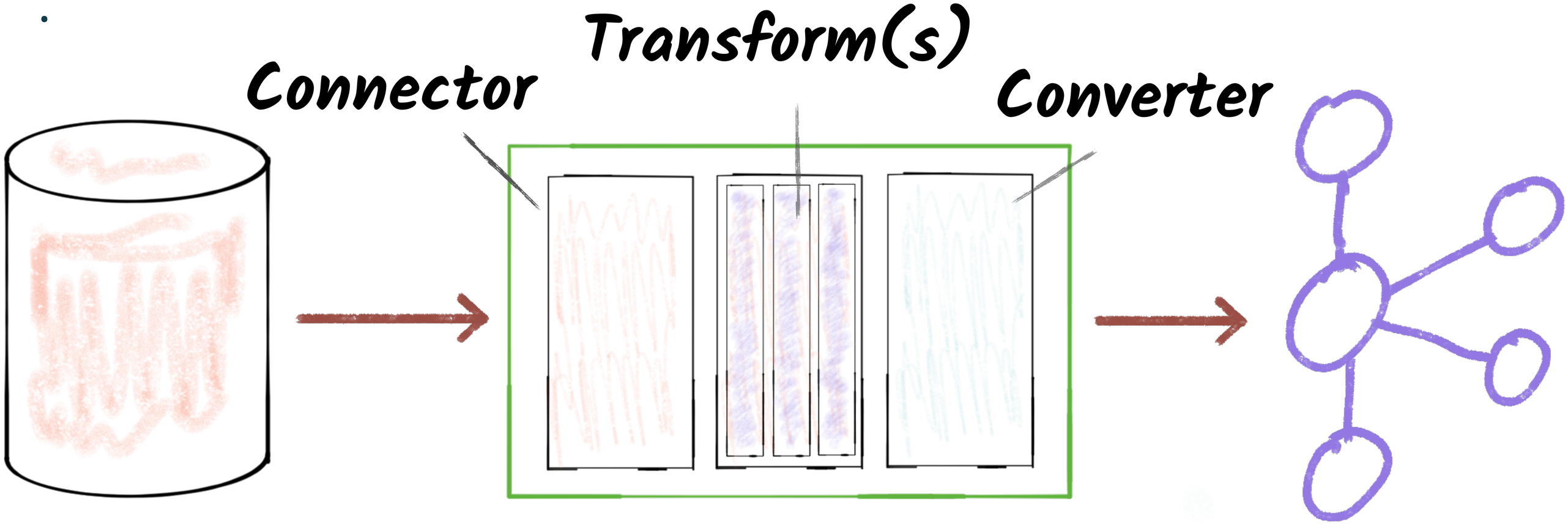
Kafka Connect

Reliable and scalable integration of Kafka with other systems – no coding required.

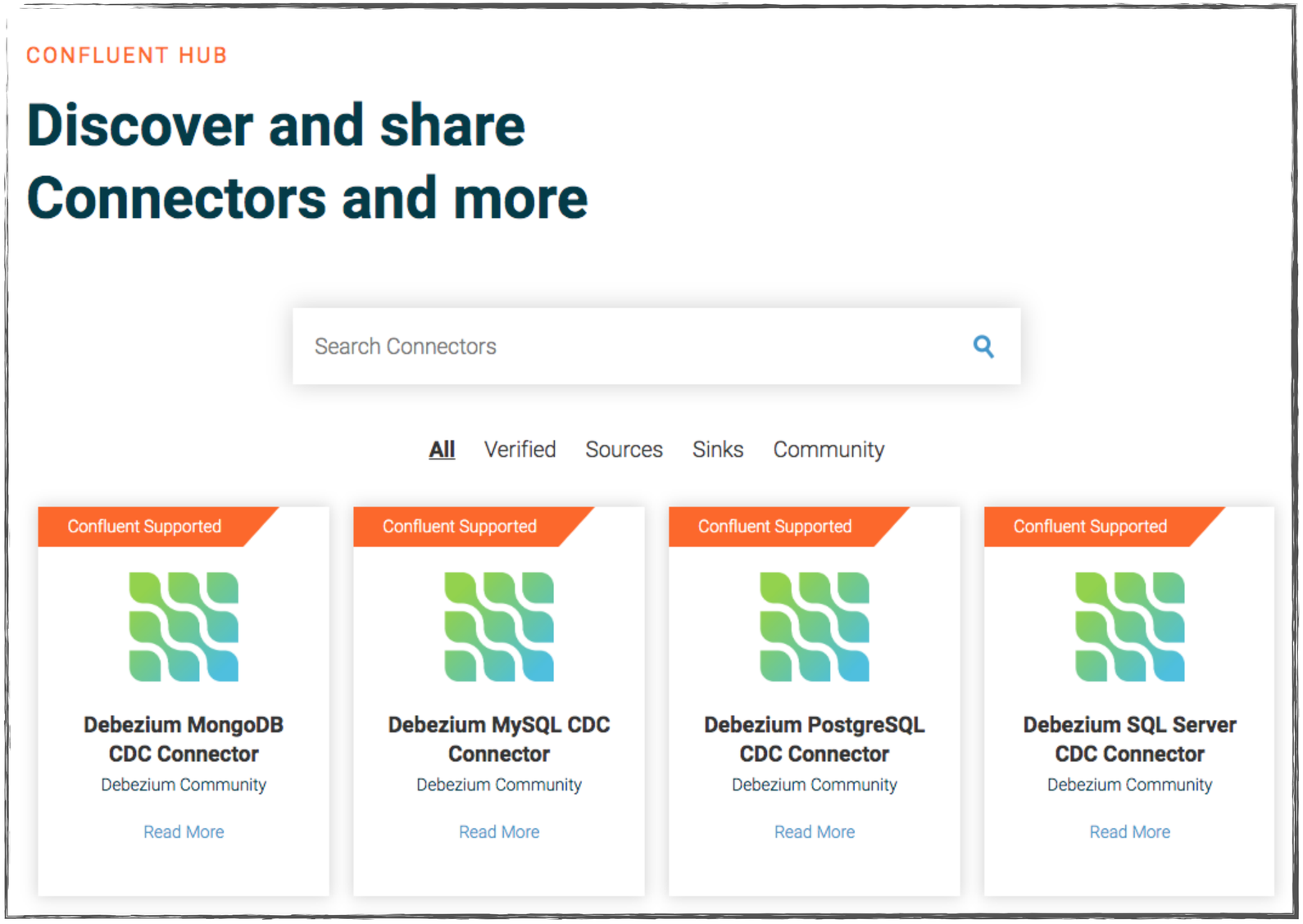
```
{
"connector.class":
    "io.confluent.connect.jdbc.JdbcSourceConnector",
"connection.url":
    "jdbc:mysql://localhost:3306/demo?user=rmoff&password=foo",
"table.whitelist":
    "sales,orders,customers"
}
```

https://docs.confluent.io/current/connect/

Extensible

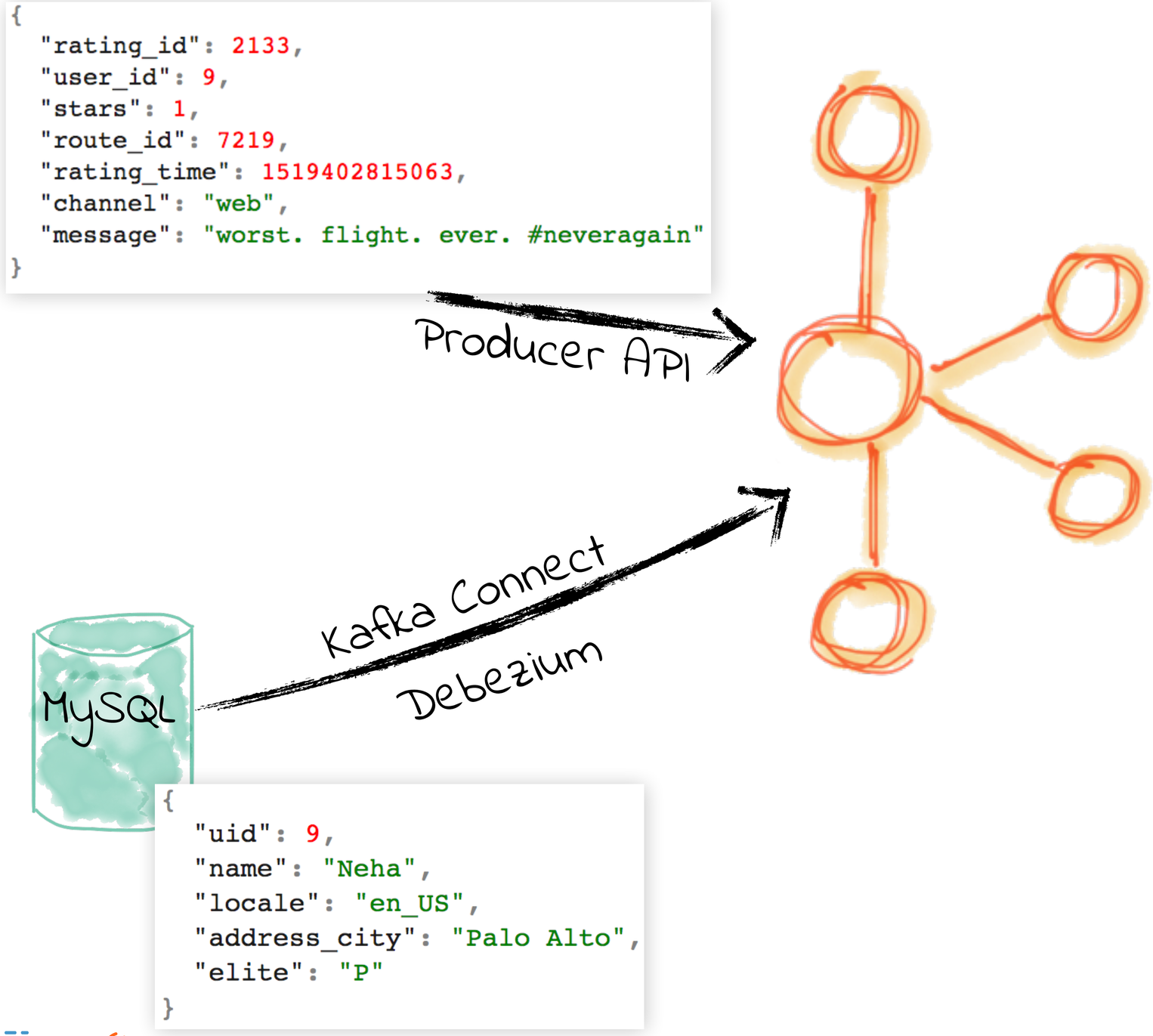


Confluent Hub



hub.confluent.io

Demo Time!


{
"rating_id": 2133,
"user_id": 9,
"stars": 1,
"route_id": 7219,
"rating_time": 1519402815063,
"channel": "web",
"message": "worst. flight. ever. #neveragain"
}
Producer API
Kafka Connect
Debezium
MySQL
{
"uid": 9,
"name": "Neha",
"locale": "en_US",
"address_city": "Palo Alto",
"elite": "P"
}

The Stream/Table Duality

Stream

Account ID	Amount
12345	+ €50
12345	+ €25
12345	-€60

Time

Table

Account ID	Balance
12345	€50

Account ID	Balance
12345	€75

Account ID	Balance
12345	€15

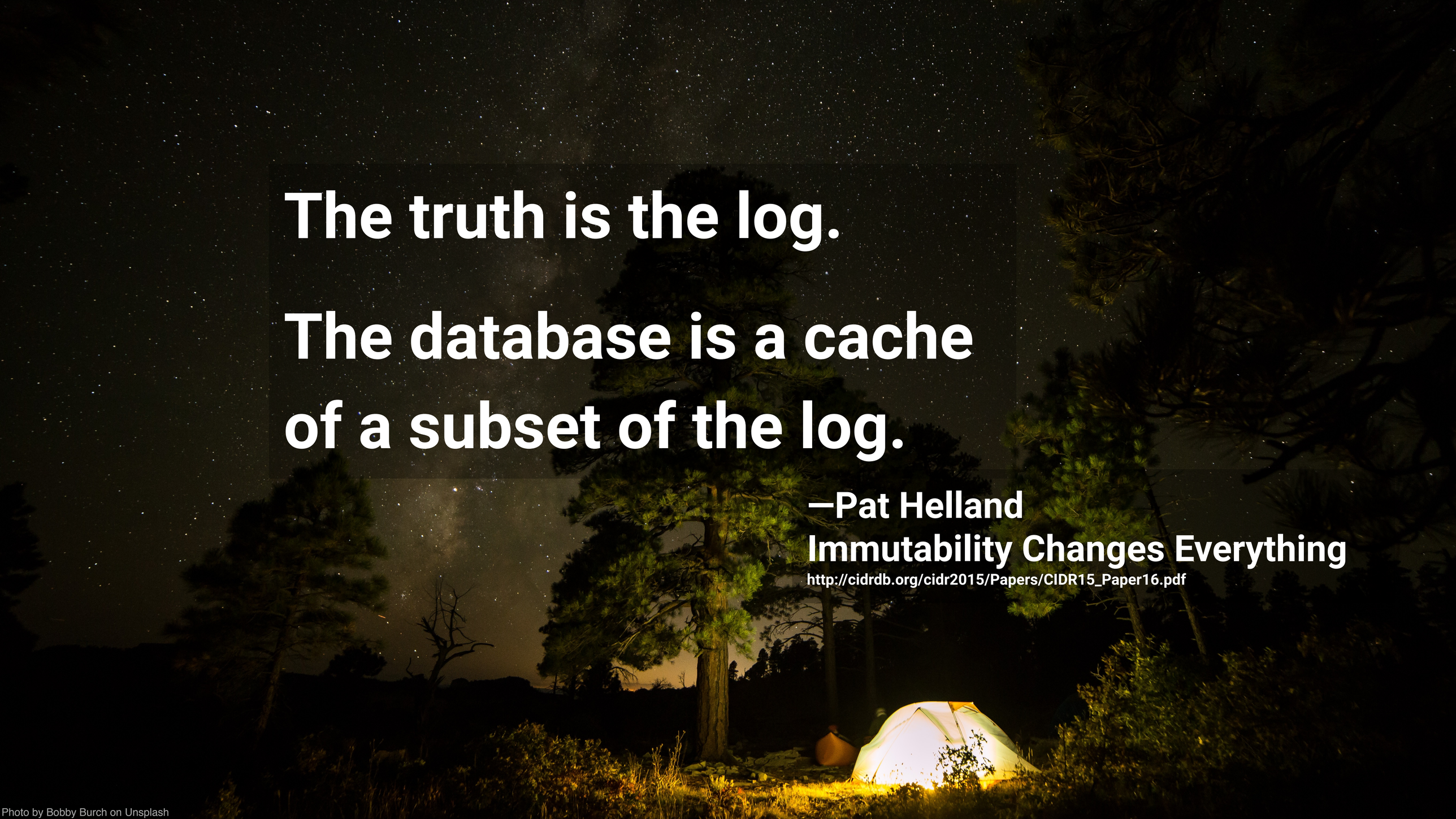
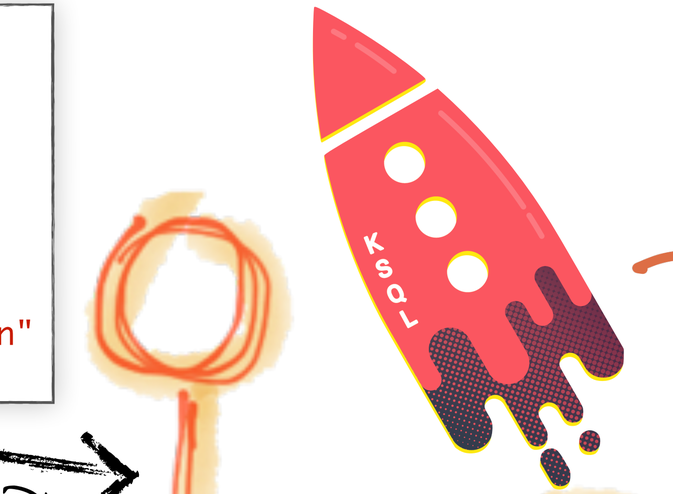


Photo by Bobby Burch on Unsplash

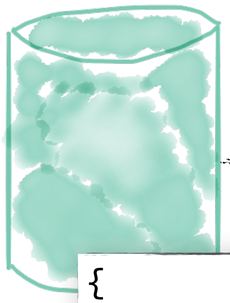
```
{
 "rating_id": 5313,
 "user_id": 3,
 "stars": 4,
 "route_id": 6975,
 "rating_time": 1519304105213,
 "channel": "web",
 "message": "worst. flight. ever. #neveragain"
}
```

Producer API



Join each rating to customer data

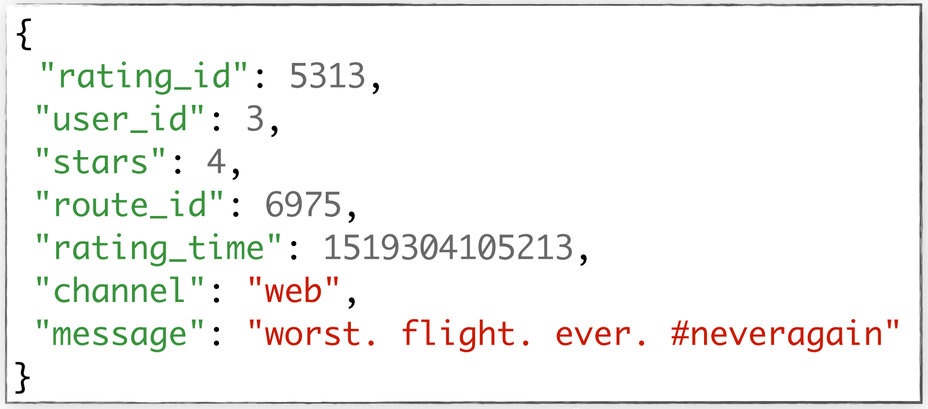
RATINGS_WITH_CUSTOMER_DATA



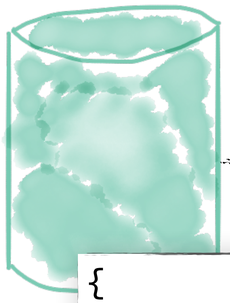
Kafka Connect

```
{
  "id": 3,
  "first_name
  "last_name"
  "email": "m
  "gender": "
  "club_statu
  "comments":
}
```

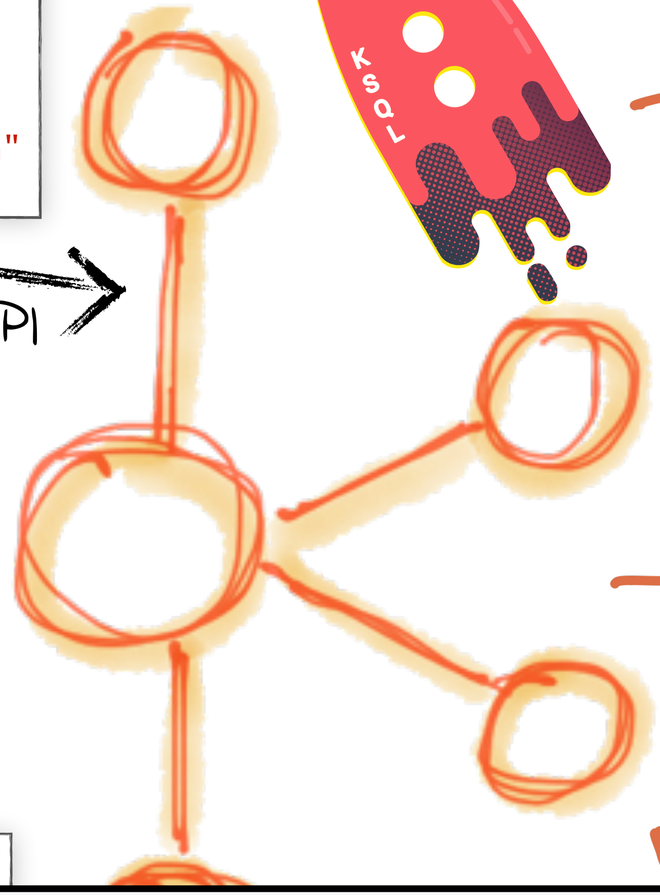
```
CREATE STREAM RATINGS_WITH_CUSTOMER_DATA AS
SELECT * FROM RATINGS
       LEFT JOIN CUSTOMERS ON R.ID=C.ID;
```



Producer API



Kafka Connect



Join each rating to customer data

RATINGS_WITH_CUSTOMER_DATA

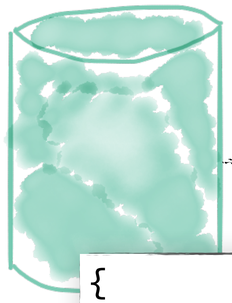
Filter for just PLATINUM customers

UNHAPPY_PLATINUM_CUSTOMERS

```
CREATE STREAM UNHAPPY_PLATINUM_CUSTOMERS AS
SELECT * FROM RATINGS_WITH_CUSTOMER_DATA
WHERE   STARS < 3
```

```
{
 "rating_id": 5313,
 "user_id": 3,
 "stars": 4,
 "route_id": 6975,
 "rating_time": 1519304
 "channel": "web",
 "message": "worst. fli
}
```


```
CREATE TABLE RATINGS_BY_CLUB_STATUS AS
SELECT      CLUB_STATUS, COUNT(*)
FROM        RATINGS_WITH_CUSTOMER_DATA
            WINDOW TUMBLING (SIZE 1 MINUTES)
GROUP BY CLUB_STATUS;
```



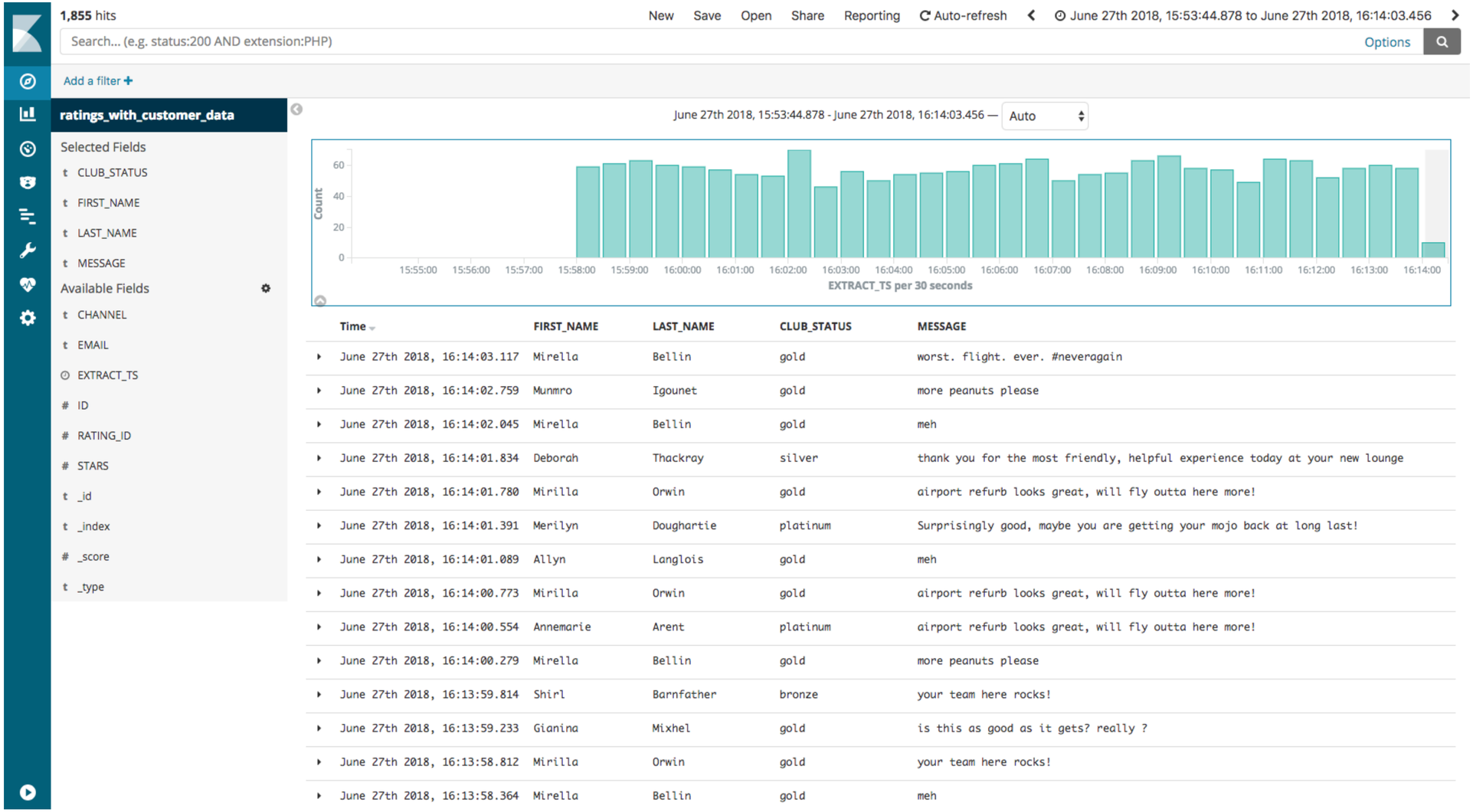
Kafka Connect

```
{
  "id": 3,
  "first_name": "Merilyn",
  "last_name": "Doughartie",
  "email": "mdoughartie1@dedecms.com",
  "gender": "Female",
  "club_status": "platinum",
  "comments": "none"
}
```


Aggregate per-minute by CLUB_STATUS

RATINGS_BY_CLUB_STATUS_1MIN

Kafka Connect → Elasticsearch



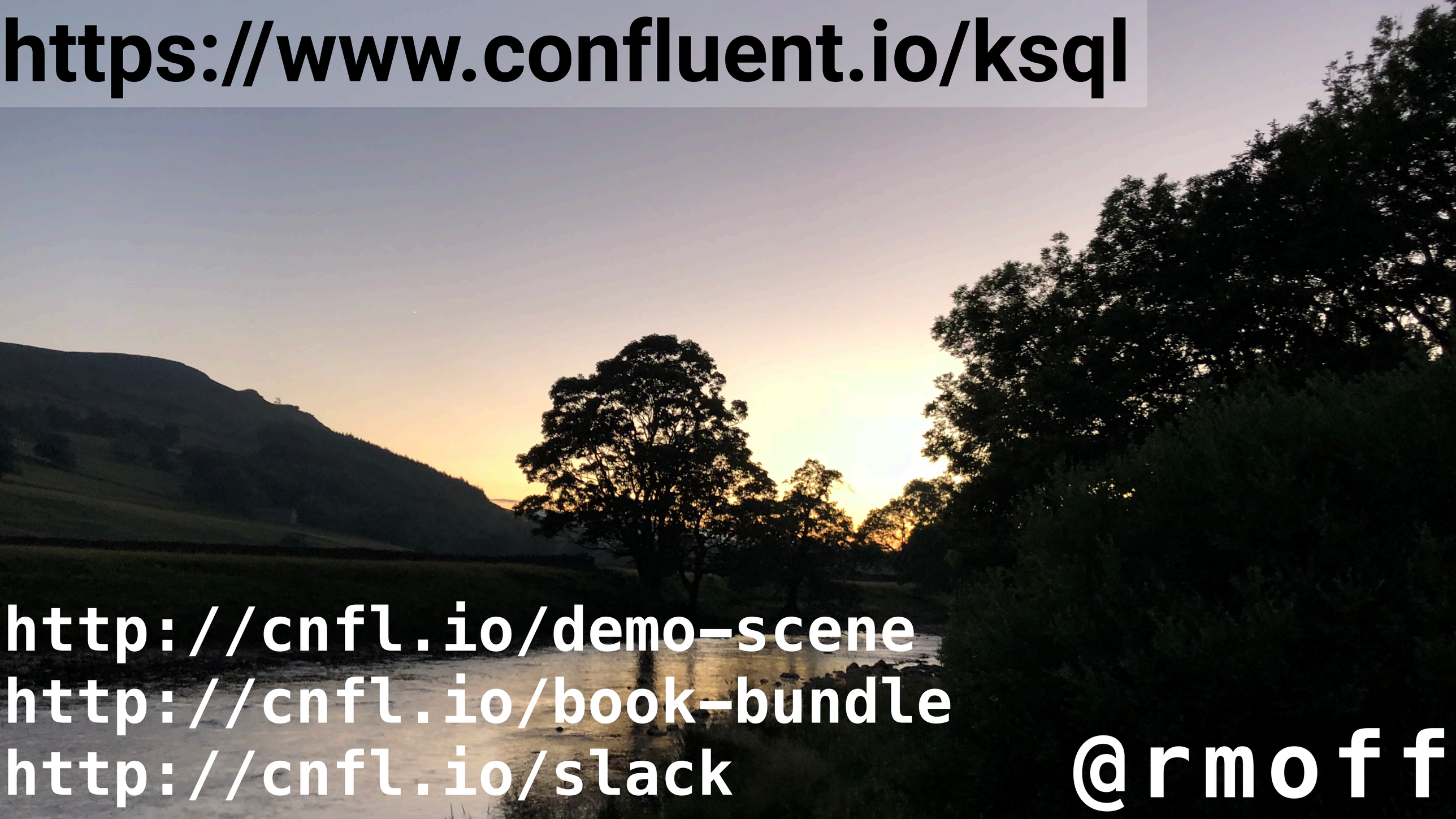
Time	FIRST_NAME	LAST_NAME	CLUB_STATUS	MESSAGE
June 27th 2018, 16:14:03.117	Mirella	Bellin	gold	worst. flight. ever. #neveragain
June 27th 2018, 16:14:02.759	Munmro	Igounet	gold	more peanuts please
June 27th 2018, 16:14:02.045	Mirella	Bellin	gold	meh
June 27th 2018, 16:14:01.834	Deborah	Thackray	silver	thank you for the most friendly, helpful experience today at your new lounge
June 27th 2018, 16:14:01.780	Mirilla	Orwin	gold	airport refurb looks great, will fly outta here more!
June 27th 2018, 16:14:01.391	Merilyn	Doughartie	platinum	Surprisingly good, maybe you are getting your mojo back at long last!
June 27th 2018, 16:14:01.089	Allyn	Langlois	gold	meh
June 27th 2018, 16:14:00.773	Mirilla	Orwin	gold	airport refurb looks great, will fly outta here more!
June 27th 2018, 16:14:00.554	Annemarie	Arent	platinum	airport refurb looks great, will fly outta here more!
June 27th 2018, 16:14:00.279	Mirella	Bellin	gold	more peanuts please
June 27th 2018, 16:13:59.814	Shirl	Barnfather	bronze	your team here rocks!
June 27th 2018, 16:13:59.233	Gianina	Mixhel	gold	is this as good as it gets? really ?
June 27th 2018, 16:13:58.812	Mirilla	Orwin	gold	your team here rocks!
June 27th 2018, 16:13:58.364	Mirella	Bellin	gold	meh

kafka summit

ORGANIZED BY confluent

Discount code!
KS19Comm25

kafka-summit.org


https://www.confluent.io/ksql
http://cnfl.io/demo-scene
http://cnfl.io/book-bundle
http://cnfl.io/slack
@rmoff

Related Talks

- The Changing Face of ETL: Event-Driven Architectures for Data Engineers
  - 📖 Slides
  - 📽️ Recording
- ATM Fraud detection with Kafka and KSQL
  - 📖 Slides
  - 👾 Code
  - 📽️ Recording
- Embrace the Anarchy: Apache Kafka's Role in Modern Data Architectures
  - 📖 Slides
  - 📽️ Recording
- Apache Kafka and KSQL in Action : Let's Build a Streaming Data Pipeline!
  - 📖 Slides
  - 👾 Code
  - 📽️ Recording
- No More Silos: Integrating Databases and Apache Kafka
  - 📖 Slides
  - 👾 Code (MySQL)
  - 👾 Code (Oracle)
  - 📽️ Recording

#EOF